考试重点：

考试难度类似

选择：不一定单选，25 分 5*5（答案不一定唯一）部分选对，没有选全，给 3 分， 部分选有错误，2 分 全选 0 分

判断：25 分 5*5（对或错）

大题：50 分

20 分 与存储有关

30 分 与计算有关 自行设计，给出解决方案 有一系列的需求，你需要解决，给出设计方案

计算类的画图说明，不需要写代码 画出计算过程就可以了

把答案写在答题纸上

重点是课后思考题

思考题会变换形式

大题一、云盘服务架构设计满足以下需求

（1）从用户的角度看，付费弹性扩展存储容量和上传/下载速度（打包成服务，按照服务来分发给用户）

（2）从服务商的角度看，弹性扩展云盘总的存储量（增加存储节点，为什么？添加改进 ring）

（3）他人分享的文件，为什么“秒到”我们的云盘（增加索引，元数据，文件记录，权限可见于不可见）

（4）恢复云盘删除的内容，付费恢复更长一段时间删除的内容（权限可见，没有删除，时间点 T1，T2）

大题二、淘宝订单数据年度统计分析

（1）自己设计订单数据结构，包括基本信息，买家卖家信息等等（多个表连接，数据太多了）

（2）选择适合的统计分析计算架构（mapreduce，wordcount）

（3）画图说明（2）满足 3 个例子，a.统计不同省份不同商品的金额 b.统计不同省份不同商品的支付方式 c.不同省份不同商品不同快递公司物流 p 天内超过 q%（线路为 key，value 是 1,1 一个统计总的，一个统计满足条件的，需要判断）

选择题: 云与服务概念， 云架构基本层次，虚拟化概念，图迭代计算，cpa 理论

判断题: 租户，kafka，批量计算和流计算的区别，分布式并行，openstack

大题: lab2 关于商品统计分类的原题，设计一个云盘(有 vip，可扩容，可找回删除数据功能等)

1.用户购买的是虚拟化的存储空间，用户体验到的是用户拥有了一定量的空间，实际上是用户用

了多少空间，服务商为他提供多少空间

2.服务商按需分配给用户空间，对大文件和重复文件建立索引，不要重复存储

3.因为文件在分享时，分享的并不是文件本身，而是文件地址的索引，别人只是吧文件的索引分享给用户了。

云端已经有人上传了相同的文件，系统只不过是在你的资料目录里添加了一条指向该文件（公共文件）的信息。

4.文件删除时其实只是删除了用户与文件之间的索引，但是文件还保存在服务器中，服务器是为了防止用户误删除而进行的保存，当一段时间后（具有时限），服务器才会删除相关文件。会员可以将文件保存在服务器上的时间更久一些。

https://blog.csdn.net/xyilu/article/details/9066973

MapReduce计算练习

- 写MapReduce的技巧
  - 将计算最终结果所需内容设置为<KEY，VALUE>组合
  - 根据实际情况分配哪些归为KEY，哪些归为VALUE
  - Map：生成<KEY，VALUE>对，分配依据为KEY或者KEY的一部分
  - Combiner：Map结果的合并，减少单个Map传递给单个Reduce的中间结果的重复
  - Partition：分发策略，reduce计算所需所有内容需要分配到一个节点
  - Reduce：计算得到最终结果

流式计算练习

- 设计拓扑结构的技巧
  - 数据生成的点
  - 数据处理的点
  - 点和点连接的边：分发策略

图数据计算

图的拆分：键值对

- ID；distance，color，weight
  - ID：顶点
  - Distance：距离，MAX表示无穷大
  - Color：着色
    - 0：白色，未被计算的节点，即未连通的节点，距离为MAX
    - 1：灰色，计算过程中节点
    - 2：黑色，已经完成最短路径计算的节点

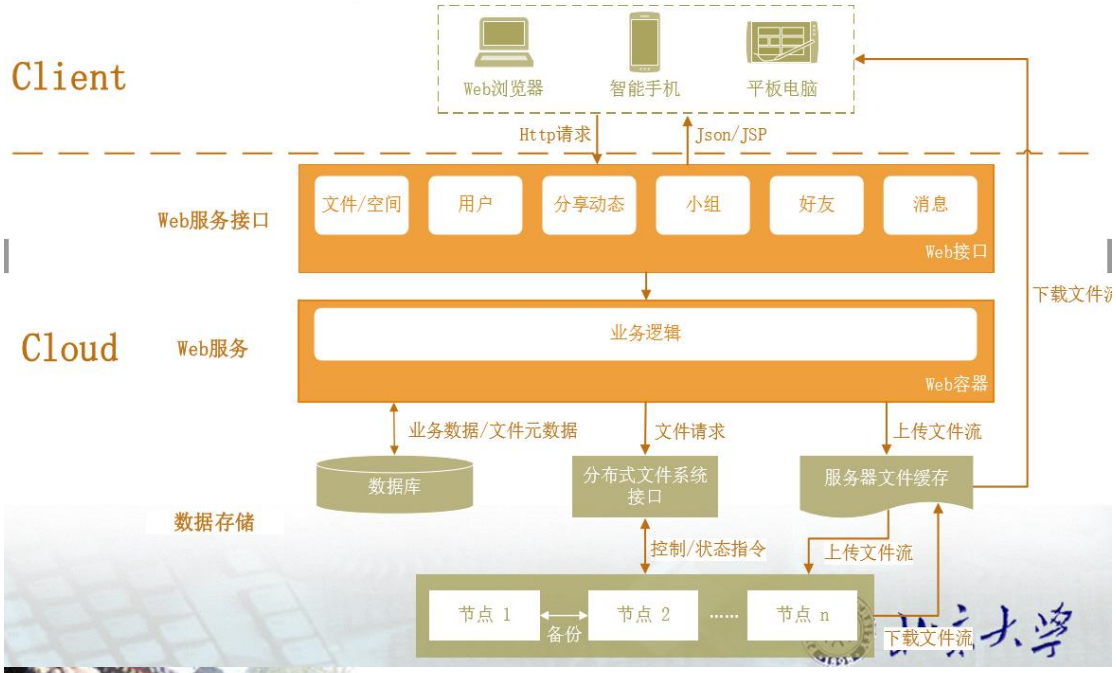
- 倒排索引算法思想示意图

Term	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	Doc 6	Doc 7	Doc 8
aid	0	0	0	1	0	0	0	1
all	0	1	0	1	0	1	0	0
back	1	0	1	0	0	0	1	0
brown	1	0	1	0	1	0	1	0
come	0	1	0	1	0	1	0	1
dog	0	0	1	0	1	0	0	0
fox	0	0	1	0	1	0	1	0
good	0	1	0	1	0	1	0	1
jump	0	0	1	0	0	0	0	0
lazy	1	0	1	0	1	0	1	0
men	0	1	0	1	0	0	0	1
now	0	1	0	0	0	1	0	1
over	1	0	1	0	1	0	1	1
party	0	0	0	0	0	1	0	1
quick	1	0	1	0	0	0	0	0
their	1	0	0	0	1	0	1	0

Term	Postings
aid	4 → 8
all	2 → 4 → 6
back	1 → 3 → 7
brown	1 → 3 → 5 → 7
come	2 → 4 → 6 → 8
dog	3 → 5
fox	3 → 5 → 7
good	2 → 4 → 6 → 8
jump	3
lazy	1 → 3 → 5 → 7
men	2 → 4 → 8
now	2 → 6 → 8
over	1 → 3 → 5 → 7 → 8
party	6 → 8
quick	1 → 3
their	1 → 5 → 7

云存储应用

- 校园网盘iStudy



云存储应用

- 文件上传流程设计：

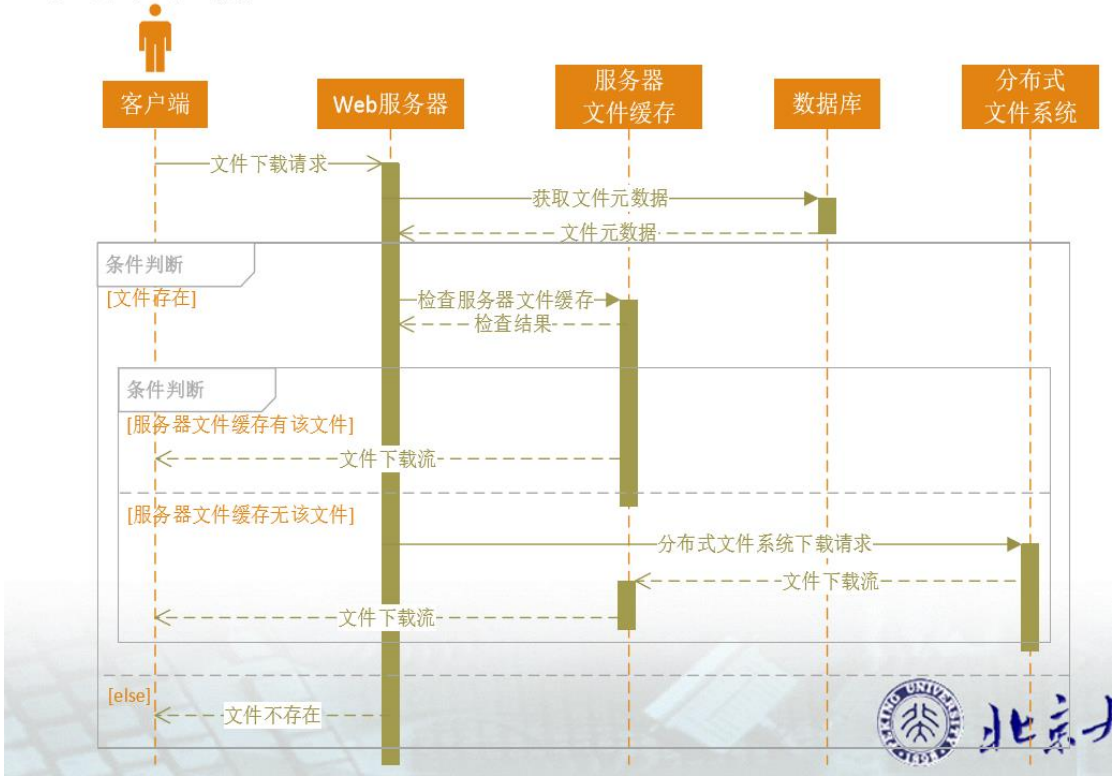
- 文件结构设计：
  - 元数据：描述文件，与真实文件一一对应，存储在数据库中

文件 = 真实文件数据 + 文件元数据（文件名，url，时间，大小，···）
(分布式文件系统) (数据库)

  - 文件的操作很容易转化为文件元数据的操作：文件分享，文件列表，添加评论、标签等

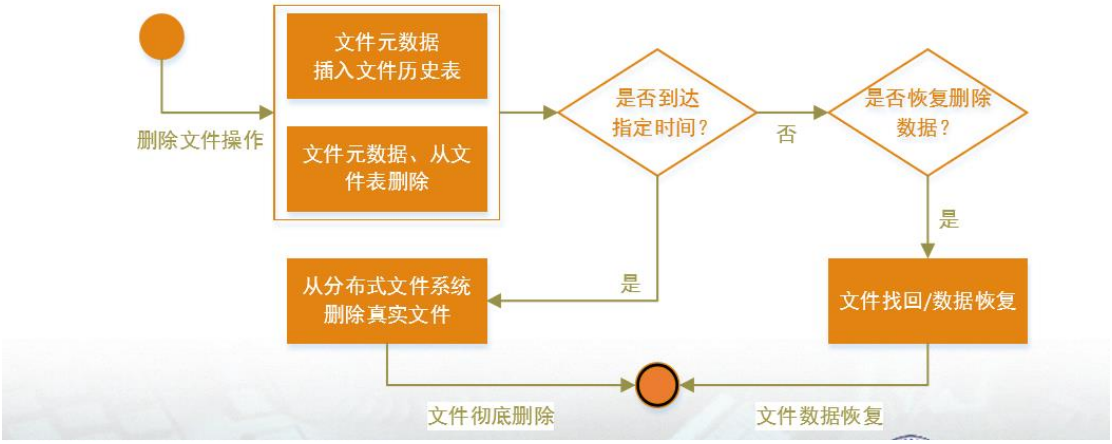
云存储

- 文件下载

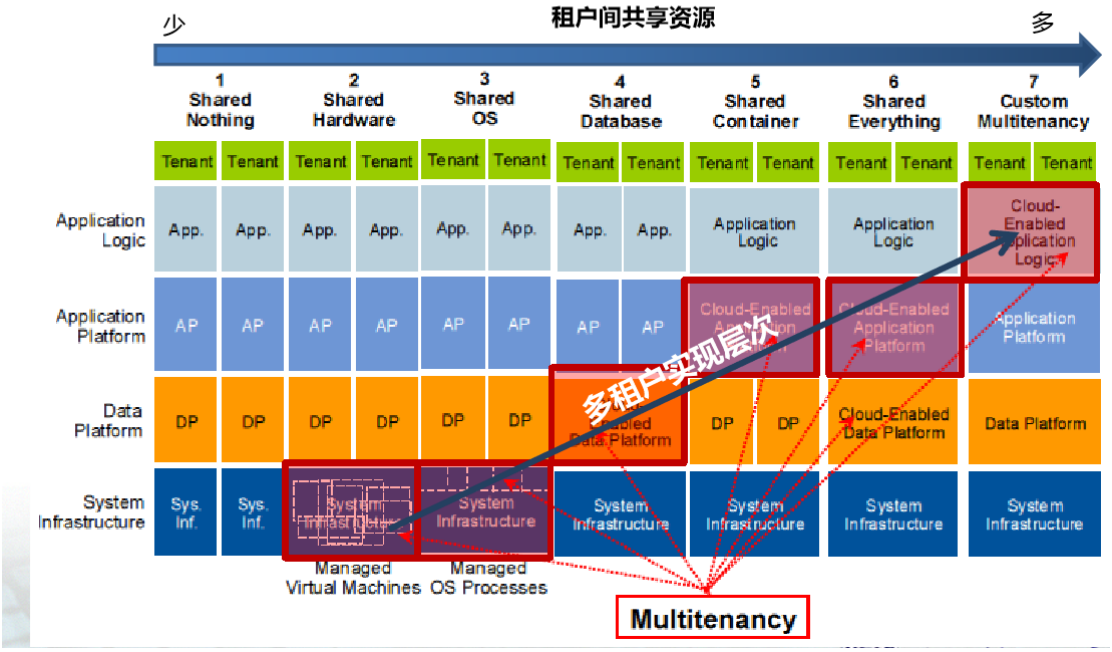


- 校园数据共享下的数据存储特点：大量重复
- 数据去重机制
  - 利用真实数据与元数据相互分离 ➡ 节省存储空间
  - 增加对元数据的引用
  - 只保存一份真实数据

- 用户删除的文件数据能否恢复？
- 利用真实数据与元数据相互分离 → 数据恢复
  - 文件延时删除



- 多租户技术



去年是 AB2 套题，所以 b 套大题是考了一个 迪杰斯特拉 算法，课堂作业题原题 去年大题基本上都是课堂作业原题或者轻微改动

By li